Comparing locality of reference - Some folk theorems for the miss rates and the output of caches

by Armand M. Makowski, Sarut Vanichpun

CSHCN TR 2004-6
(ISR TR 2004-10)

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2004**	2. REPORT TYPE	3. DATES COVERED **00-00-2004 to 00-00-2004**

4. TITLE AND SUBTITLE **Comparing Locality of Reference - Some Folk Theorems for the Miss Rates and the Output of Caches**	5a. CONTRACT NUMBER 5b. GRANT NUMBER 5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER 5e. TASK NUMBER 5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S) 11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**34**	

Chapter 1

COMPARING LOCALITY OF REFERENCE – SOME FOLK THEOREMS FOR THE MISS RATES AND THE OUTPUT OF CACHES

Armand M. Makowski
University of Maryland
Corresponding author
armand@isr.umd.edu

Sarut Vanichpun
University of Maryland
sarut@eng.umd.edu

Abstract The performance of demand-driven caching is known to depend on the locality of reference exhibited by the stream of requests made to the cache. In spite of numerous efforts, no consensus has been reached on how to formalize this notion, let alone on how to compare streams of requests on the basis of their locality of reference. We take on this issue with an eye towards validating operational expectations associated with the notion of locality of reference. We focus on two "folk theorems," namely (i) The stronger locality of reference, the smaller the miss rate of the cache; (ii) Good caching is expected to produce an output stream of requests exhibiting less locality of reference than the input stream of requests.

We discuss these two folk theorems in the context of a cache operating under a demand-driven replacement policy when document requests are modeled according to the Independent Reference Model (IRM). As we propose to measure strength of locality of reference in a stream of requests through the skewness of its popularity distribution, we introduce the notion of majorization as a mean for capturing this degree of skewness. We show that these folk theorems hold for caches operating under a large class of cache replacement policies, including the optimal policy A_0 and the random policy, but may fail under the LRU policy.

1. Introduction

Web caching aims to reduce network traffic, server load and user-perceived retrieval latency by replicating "popular" content on (proxy) caches that are strategically placed within the network, e.g., Wang (1999) (and references therein). This approach is a natural outgrowth of caching techniques which were originally developed for computer memory and distributed file sharing systems, e.g., Aven, Coffman and Kogan (1987); Coffman and Denning (1973); Phalke and Gopinath (1995) (and references therein). However, the exponential growth of the World Wide Web and its specific circumstances are challenging current cache architectures to meet the complementary mandates of speed, scalability and reliability which are central to delivering a satisfactory user experience.

Although these challenges have renewed interest in caching in general, some basic issues are still not well understood. Indeed, the performance of any form of caching is determined by a number of factors, chief amongst them the statistical properties of the streams of requests made to the cache. One important such property is the *locality of reference* present in a stream of requests whereby "bursts of references are made in the near future to objects referenced in the recent past." The importance of locality for caching was first recognized by Belady (1966) in the context of computer memory, and attempts at characterizing it were made early on by Denning (1968) through the working set model. Recently, a number of studies have shown that streams of requests for Web objects exhibit strong locality of reference[1] (see e.g., Jin and Bestavros (2000b); Mahanti, Williamson and Eager (2000)). Like the notion of burstiness used in traffic modeling, locality of reference, while endowed with a clear intuitive content, admits no simple definition.

Thus, and not surprisingly, in spite of numerous efforts, no consensus has been reached on how to formalize the notion, let alone *compare* streams of requests on the basis of their locality of reference.[2] To the best of the authors' knowledge, this lack of consensus has precluded the formal derivation of the following "folk theorems":

1. **Folk theorem on miss rates** – The stronger the locality of reference in the stream of requests, the smaller the miss rate since the cache ends up being populated by objects with a higher likelihood of access in the near future. Such a property, if true, would confirm the central role played by locality of reference in shaping cache

[1] At least in the short timescales

[2] An exception can be found in a recent paper by Fonseca et al. (2003).

performance. In fact, the very presence of locality of reference in the stream of requests is what makes caching at all possible; and

2. **Folk theorem on output streams** – Good cache replacement strategies "absorb" locality of reference to a certain extent by producing a stream of misses from the cache – its so-called output – which exhibits *less* locality of reference than the input stream of requests. In the context of multi-level caching, this reduction property is often perceived as one of the main reasons for why caching looses its effectiveness after some level in a hierarchy of caches.

Such folk theorems are expected to hold for demand-driven caching that exploits recency of reference. Interest in establishing them under a *specific* definition of locality of reference stems from a desire to validate its *operational* significance. Counterexamples would cast some doubts as to whether the particular definition indeed captures the intuitive meaning of locality of reference.

In the past such a program has been carried out for a number of key notions of traffic engineering: For instance, the convex stochastic orderings were shown to capture the notion of *variability*, in the process leading to various proofs that "determinism minimizes waiting times," e.g., Baccelli and Makowski (1989). More recently, the theory of multivariate stochastic orderings has been used to formalize the belief that *positive correlations* lead to larger buffer levels at a discrete-time infinite capacity multiplexer queue, viz. if the input traffic is larger than its independent version in the supermodular ordering, then their corresponding buffer contents are similarly ordered in the increasing convex ordering. This has been demonstrated for a number of basic traffic models in Vanichpun and Makowski (2002).

In this chapter we survey and extend recent results by the authors concerning a formal investigation into the folk theorems mentioned earlier, albeit in a simple framework. The results for miss rates and output streams are available in Vanichpun and Makowski (2004a) and Vanichpun and Makowski (2004b), respectively, and the interested reader is referred to these papers for additional information. In the next section, we provide a roadmap to the viewpoint we have adopted and to the ensuing results, as well as the organization of the chapter.

2. Navigating the chapter - A roadmap

2.1 Locality via popularity

Our first task consists in identifying the notion of locality of reference to be used here. We begin with the widely accepted observation that the two main contributors to locality of reference are *temporal correlations* in the streams of requests and the *popularity distribution* of requested objects. To describe these two sources of locality, we assume the following generic setup which is used throughout: We consider N cacheable items or documents, labeled $i = 1, \ldots, N$, and we write $\mathcal{N} := \{1, \ldots, N\}$. The successive requests arriving at the cache are modeled by a sequence $\{R_t,\ t = 0, 1, \ldots\}$ of $\mathcal{N}$-valued rvs.

1. The *popularity* of the sequence of requests $\{R_t,\ t = 0, 1, \ldots\}$ is defined as the pmf $\boldsymbol{p} = (p(i), \ldots, p(N))$ on $\mathcal{N}$ given by

$$p(i) := \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbf{1}\left[R_\tau = i\right] \quad a.s., \quad i = 1, \ldots, N \tag{1.1}$$

whenever these limits exist (and they do in most models treated in the literature).

2. *Temporal correlations* are more delicate to define due to the "categorical" nature of the requests $\{R_t,\ t = 0, 1, \ldots\}$. Indeed, it is somewhat meaningless to use the covariance function

$$\gamma(s, t) := \mathrm{Cov}[R_s, R_t], \quad s, t = 0, 1, \ldots.$$

as a way to capture these temporal correlations as is traditionally done in other contexts. This is because the rvs $\{R_t,\ t = 0, 1, \ldots\}$ take values in a discrete set. We took $\{1, \ldots, N\}$ but could have selected *any* set of N distinct points in an arbitrary space. Thus, the *actual* values of the rvs $\{R_t,\ t = 0, 1, \ldots\}$ are of no consequence, and the focus should instead be on the *recurrence patterns* exhibited by requests for particular documents over time. The literature contains several metrics to do this, including the inter-reference time of Phalke and Gopinath (1995), the working set size of Denning (1968) and the stack distance, see e.g., Almeida et al. (1996).

We shall focus *exclusively* on popularity as the measure of locality of reference. In fact, to isolate its contribution, we deal with the situation where there is *no* temporal correlations in the stream of requests as would be the case under the so-called *Independence Reference Model* (IRM). More precisely, under the IRM with popularity pmf $\boldsymbol{p} = (p(1), \ldots, p(N))$, the successive requests $\{R_t,\ t = 0, 1, \ldots\}$ form a sequence of i.i.d. $\mathcal{N}$-valued rvs, each distributed according to the pmf $\boldsymbol{p}$, i.e.,

$$\mathbf{P}\left[R_t = i\right] = p(i), \quad i = 1, \ldots, N \tag{1.2}$$

for all $t = 0, 1, \ldots$ and (1.1) holds with the given pmf $\boldsymbol{p}$ by the Law of Large Numbers.

IRMs *do* display locality of reference even though there is no temporal correlations. This is best appreciated by considering the limiting cases: If $\boldsymbol{p}$ is extremely unbalanced with $\boldsymbol{p} = (1-\delta, \varepsilon, \ldots, \varepsilon)$ (with $\delta = (N-1)\varepsilon$), a reference to document 1 is likely to be followed by a burst of additional references to document 1 provided $(N-1)\varepsilon \ll 1-\delta$. It seems natural to deem this situation as one exhibiting very strong locality of reference. The exact opposite conclusion holds if the popularity pmf $\boldsymbol{p}$ were uniform, i.e., $p(1) = \ldots = p(N) = \frac{1}{N}$, for then the successive requests $\{R_t,\ t = 0, 1, \ldots\}$ form a truly random sequence, in which case there is no locality of reference. Thus, the *skewness* of $\boldsymbol{p}$ appears to act as an indicator of the strength of locality of reference present in the stream, under the intuition that the more "balanced" the pmf $\boldsymbol{p}$, the weaker the locality of reference.

2.2 Majorization and Schur-concavity

As we restrict ourselves to the class of IRMs,[3] the question naturally arises as to whether popularity pmfs can be compared on the basis of their skewness so that versions of the folk theorems discussed earlier can be established. More formally, consider two IRMs with popularity pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ (on $\mathcal{N}$), and let $M(\boldsymbol{p})$ and $M(\boldsymbol{q})$ denote their miss rates under some cache replacement policy. We seek a way to *formally* compare the pmf vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, with the interpretation that if $\boldsymbol{p}$ is less skewed than $\boldsymbol{q}$, then the IRM with popularity pmf $\boldsymbol{p}$ has less locality of reference than the IRM with popularity pmf $\boldsymbol{q}$, and the folk theorem on miss rates holds as

$$M(\boldsymbol{q}) \leq M(\boldsymbol{p}). \tag{1.3}$$

We turn to the concept of *majorization* discussed in the monograph of Marshall and Olkin (1979) as a way to characterize such imbalance in the components of popularity pmfs. Motivated by our earlier discussion, we say that the IRM with popularity pmf $\boldsymbol{p}$ has less locality of reference than the IRM with popularity pmf $\boldsymbol{q}$ if $\boldsymbol{p}$ is *majorized* by $\boldsymbol{q}$, written $\boldsymbol{p} \prec \boldsymbol{q}$. As elegantly demonstrated in the monograph of Marshall and Olkin (1979), this notion has found widespread use in many diverse branches of mathematics and their applications. What is more, comparison results such as (1.3) can now be explored through the rich and structured class of monotone functions associated with majorization, the so-called Schur-

[3] This may not be too much of a limitation given that the IRM is the most basic request model; it is often used for checking various properties, see e.g., Breslau et al. (1999). Moreover, recent results by Jelenkovic and Radovanovic (2003) suggest some form of insensitivity to the statistics of streams of requests. Of course, more work along these lines is needed.

convex/concave functions. In fact, the comparison (1.3) is essentially a statement concerning the Schur-concavity of certain functionals.

Within this framework, if $\boldsymbol{p}^\star$ denotes the popularity pmf for the output from the cache, then the folk theorem on the stream of misses takes the form

$$\boldsymbol{p}^\star \prec \boldsymbol{p}. \tag{1.4}$$

Both statements (1.3) and (1.4) were investigated in the context of a cache operating under a demand-driven replacement policy when document requests are modeled according to the IRM. We now summarize some of the findings.

2.3 The folk theorems under RORA policies

In Vanichpun and Makowski (2004a) and Vanichpun and Makowski (2004b), the authors have shown the validity of both statements (1.3) and (1.4) for a number of policies, namely the optimal policy A_0, the random policy and the First-In/First-Out (FIFO) policy. These properties hold in *all* circumstances, i.e., for an *arbitrary* popularity pmf for the IRM input and for *arbitrary* cache sizes. To the best of the authors' knowledge, these results provide the first formal proof of the folk theorems. In this chapter, we have extended these positive results to a very large class of replacement policies, known as Random On-demand Replacement Algorithms (RORA); these policies generalize the policy A_0, the random policy and the FIFO policy.

2.4 Counterexamples and asymptotics

However, there are policies for which the comparisons (1.3) and (1.4) do *not* always hold. One such policy is the *Least-Recently-Used* (LRU) replacement policy, a popular self-organizing eviction policy. Indeed, we first exhibit situations where the miss rate of the LRU policy is larger when selecting an IRM with a more balanced popularity pmf. Yet, when the popularity pmfs are Zipf-like, simulations show that the comparison (1.3) still *does* hold for the LRU policy. We formally establish this fact only in the limiting regime where the skewness parameter of the Zipf-like pmf is large, i.e., highly skewed.

It also happens that the LRU policy fails to reduce locality of reference in the sense of (1.4). We explore the issue through counterexamples which are developed within the class of Zipf-like popularity pmfs. For this class of input pmfs, we identify a condition involving the cache size and the number of cacheable documents under which (1.4) fails to occur at large enough values of the skewness parameter of the Zipf-like pmf. Under this condition, which is reasonably satisfied in practice, we show that the output pmf $\boldsymbol{p}^\star$ may not exhibit less locality of reference than the input pmf $\boldsymbol{p}$ when the latter has too much of it to begin with. Additional simulations were carried out and suggested a conjecture as to when LRU caching indeed reduces locality of reference with Zipf-like input pmfs. All indications point to the possibility that for small enough cache sizes, the desired comparison of $\boldsymbol{p}$ and $\boldsymbol{p}^\star$ will hold; this will be the subject of future investigation.

While the discussion given here is restricted to IRMs, we believe that similar results may hold for more general input models.

2.5 Organization

The chapter is organized as follows: The basic model of cache management is given in Section 1.3. The miss rate and output of a cache are discussed in Section 1.4 and 1.5, respectively. Majorization and the companion notion of Schur-convexity are introduced in Section 1.6 and 1.7, respectively. We obtain the basic comparison results for the output in Section 1.8. The RORA cache policies are defined in Section 1.9, and the comparison results for their miss rates and outputs are given in Section 1.10 and 1.11, respectively. Zipf-like distributions are discussed in Section 1.12. Comparison results for the miss rate and output under the LRU policy are collected in Section 1.13 and 1.14, respectively.

3. Demand-driven caching

Consider a universe $\mathcal{N} = \{1, \ldots, N\}$ of N cacheable documents. The system is composed of a server where a copy of each of these N documents is available, and of a cache of size M $(1 \leq M < N)$. Documents are first requested at the cache: If the requested document has a copy already in cache (i.e., a hit), this copy is downloaded from the cache by the user. If the requested document is not in cache (i.e., a miss), a copy is requested instead from the server to be put in the cache. If the cache is already full, then a document already in cache is evicted to make place for the copy of the document just requested. The document selected for eviction is determined through a *cache replacement* or *eviction* policy.[4]

We now develop below a mathematical framework to address some of the issues discussed in this chapter. Additional details are available in the monographs by Aven, Coffman and Kogan (1987) and by Coffman and Denning (1973). We begin with some notation that will be used repeatedly: Let $\Lambda^\star(M;\mathcal{N})$ be the collection of all *unordered* subsets of size M of $\mathcal{N}$, and let $\Lambda(M;\mathcal{N})$ be the collection of all *ordered* sequences of

[4] We use the terms interchangeably.

M *distinct* elements from $\mathcal{N}$. We write $\{i_1, \ldots, i_M\}$ (resp. $(i_1, \ldots, i_M)$) to denote an element in $\Lambda^\star(M; \mathcal{N})$ (resp. $\Lambda(M; \mathcal{N})$).

3.1 A simple framework

Consecutive user requests are modeled by a sequence of $\mathcal{N}$-valued rvs $\{R_t,\ t = 0, 1, \ldots\}$. For simplicity we say that request R_t occurs at time $t = 0, 1, \ldots$. Let S_t denote the cache just before time t so that S_t is a subset of $\mathcal{N}$ with at most M elements. The decision to be performed according to the eviction policy in force is the identity U_t of the document in S_t which needs to be evicted in order to make room for the request R_t (if the cache is already full).

Demand-driven caching considered here is characterized by the dynamics

$$S_{t+1} \;=\; \begin{cases} S_t & \text{if } R_t \in S_t \\ S_t + R_t & \text{if } R_t \notin S_t, |S_t| < M \\ S_t - U_t + R_t & \text{if } R_t \notin S_t, |S_t| = M \end{cases} \tag{1.5}$$

for all $t = 0, 1, \ldots$, where $|S_t|$ denotes the cardinality of the set S_t, and $S_t - U_t + R_t$ denotes the subset of $\{1, \ldots, N\}$ obtained from S_t by removing U_t and then adding R_t to it, *in that order*. These dynamics reflect the following operational assumptions: (i) actions are taken only at the time requests are made, hence the expression demand-driven caching; (ii) a requested document not in cache is *always* added to the cache if the cache is not full at the time of request; and (iii) eviction is *mandatory* if the request R_t is not in cache S_t and the cache S_t is full, i.e., $|S_t| = M$.

3.2 Admissible IRMs and reduced dynamics

Throughout the stream of requests $\{R_t,\ t = 0, 1, \ldots\}$ is modeled according to the standard *Independence Reference Model* (IRM) with popularity pmf $\boldsymbol{p} = (p(1), \ldots, p(N))$. To avoid uninteresting situations, it is *always* the case that

$$p(i) > 0, \quad i = 1, \ldots, N. \tag{1.6}$$

A pmf $\boldsymbol{p}$ on $\{1, \ldots, N\}$ satisfying (1.6) is said to be *admissible*.

Under this non-triviality condition (1.6), every document will eventually be requested by virtue of (1.1). Thus, as we have in mind to study long term characteristics under demand-driven replacement policies, there is no loss of generality in assuming (as we do from now on) that the cache is full, i.e., for all $t = 0, 1, \ldots$, we have $|S_t| = M$ and (1.5) simplifies to

$$S_{t+1} \quad = \quad \begin{cases} S_t & \text{if } R_t \in S_t \\ S_t - U_t + R_t & \text{if } R_t \notin S_t. \end{cases} \tag{1.7}$$

3.3 Cache states and eviction policies

The decisions $\{U_t,\ t = 0, 1, \ldots\}$ are determined through an eviction policy and several examples will be presented shortly.

Consider a given eviction policy π. We assume that the dynamics of the cache can be characterized through the evolution of suitably defined variables $\{\Omega_t,\ t = 0, 1, \ldots\}$ where Ω_t is known as the *state of the cache* at time t. The cache state is specific to the eviction policy and is selected with the following in mind: (i) The set S_t of documents in the cache at time t can be recovered from Ω_t; (ii) the cache state Ω_{t+1} is fully determined through the knowledge of the triple (Ω_t, R_t, U_t) in a way that is compatible with the dynamics (1.7); and (iii) the eviction decision U_t at time t can be expressed as a function of the past $(\Omega_0, R_0, U_0, \ldots, \Omega_{t-1}, R_{t-1}, U_{t-1}, \Omega_t, R_t)$ (possibly through suitable randomization), i.e., for each $t = 0, 1, \ldots,$ there exists a mapping π_t such that

$$U_t = \pi_t(\Omega_0, R_0, U_0, \ldots, \Omega_{t-1}, R_{t-1}, U_{t-1}, \Omega_t, R_t; \Xi_t)$$

where the rv Ξ_t is taken independent of the past $(\Omega_0, R_0, \ldots, U_{t-1}, \Omega_t, R_t)$. Collectively the mappings $\{\pi_t,\ t = 0, 1, \ldots\}$ define the eviction policy π.

We close this section with some examples of eviction policies which have been discussed in the literature, see e.g., the monographs by Aven, Coffman and Kogan (1987) and by Coffman and Denning (1973):

According to the *random policy*, when the cache is full, the document to be evicted is selected randomly from the cache according to the uniform distribution.

Any permutation σ of $\{1, \ldots, N\}$ induces an ordering of the documents by considering the documents $\sigma(1), \sigma(2), \ldots, \sigma(N)$ as "ranked" in decreasing order. This ranking of the documents allows us to define the eviction policy A_σ as follows: When at time $t = 0, 1, \ldots,$ the cache S_t is full and the requested document R_t is not in the cache, the policy A_σ prescribes the eviction of the document U_t given by $U_t = \arg\max\left(\sigma^{-1}(j) :\ j \in S_t\right)$. The documents $\sigma(1), \ldots, \sigma(M-1)$, once loaded in the cache, will remain there, and in the steady state, the cache under the policy A_σ will contain the documents $\sigma(1), \ldots, \sigma(M-1)$.

The so-called *policy* A_0 is associated with the underlying popularity pmf $\boldsymbol{p}$ of the request stream, and evicts the least popular document in the cache, i.e., $U_t = \arg\min\left(p(j) :\ j \in S_t\right)$ for each $t = 0, 1, \ldots$. This policy A_0 coincides with the policy $A_{\sigma^\star}$ associated with the permutation

$\sigma^\star$ of $\{1, \ldots, N\}$ which orders the components of the underlying pmf $\boldsymbol{p}$ in decreasing order, namely $p(\sigma^\star(1)) \geq p(\sigma^\star(2)) \geq \ldots \geq p(\sigma^\star(N))$.

Under the random policy and the policies A_σ, we can take the cache state to be the (unordered) set of documents in the cache, i.e., the cache state is an element of $\Lambda^\star(M; \mathcal{N})$ and $\Omega_t = S_t$ for all $t = 0, 1, \ldots$.

The FIFO policy replaces the document which has been in cache for the longest time, while the LRU policy evicts the least recently requested document already in cache. The definitions of the FIFO and LRU policies necessitate that the cache state be an element of $\Lambda(M; \mathcal{N})$ with Ω_t being a permutation of the elements in S_t for all $t = 0, 1, \ldots$.

4. The miss rate of a cache

A standard performance metric to compare various caching policies is the *miss rate* of the cache. This quantity has the interpretation of being the long-term frequency of the event that the requested document is not in the cache, and therefore determines the effectiveness of a caching policy.

Under a cache replacement policy π, the miss rate $M_\pi(\boldsymbol{p})$ is defined as the a.s. limit

$$M_\pi(\boldsymbol{p}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[R_\tau \notin S_\tau\right] \quad a.s. \tag{1.8}$$

where S_τ denotes the set of documents in cache operating under the replacement policy π at time τ when the input to the cache is the request stream $\{R_t, t = 0, 1 \ldots\}$. Almost sure convergence in (1.8) (and elsewhere) is taken under the probability measure on the sequence of rvs $\{\Omega_t, R_t, U_t,\ t = 0, 1, \ldots\}$ induced by the underlying IRM with popularity pmf $\boldsymbol{p}$ through the eviction policy π.

Under most cache replacement policies of interest, the limit (1.8) exists and admits a simple expression under the assumption that the a.s. limit

$$Q_\pi(s; \boldsymbol{p}) = \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \quad a.s. \tag{1.9}$$

exists for each element s in $\Lambda^\star(M; \mathcal{N})$. Although the limits (1.8) and (1.9) are often constants which are independent of the initial cache state Ω_0, this is not always the case as be seen in the discussion of RORA policies in Sections 1.9 and 1.10.

THEOREM 1.1 *Consider an eviction policy π such that the limits (1.9) exist under the IRM with popularity pmf $\boldsymbol{p}$. Then, the limit (1.8) exists* *and is given by*

$$M_\pi(\boldsymbol{p}) = \sum_{i=1}^{N} p(i) \sum_{s\in\Lambda_i^\star(M;\mathcal{N})} Q_\pi(s;\boldsymbol{p}) \quad (1.10)$$

$$= \sum_{s\in\Lambda^\star(M;\mathcal{N})} Q_\pi(s;\boldsymbol{p}) \sum_{i\notin s} p(i) \quad (1.11)$$

where $\Lambda_i^\star(M;\mathcal{N})$ *denotes the set of elements in* $\Lambda^\star(M;\mathcal{N})$ *which do not contain* i*, i.e.,* $\Lambda_i^\star(M;\mathcal{N}) := \{s = \{i_1, \ldots i_M\} \in \Lambda^\star(M;\mathcal{N}) : \ i \notin s\}$ *.*

Theorem 1.1 is a standard result under IRMs; its proof can also be found in Vanichpun (2004). The existence of the limits (1.9) is a mild assumption which is satisfied under all eviction policies of interest considered here (and in the literature). Indeed, under the IRM with popularity pmf $\boldsymbol{p}$, the sequence of cache states $\{\Omega_t, \ t = 0, 1, \ldots\}$ typically form a Markov chain over a finite state space, and standard ergodic results readily yield the existence of the limits (1.9). This issue will be briefly discussed in each situation at the appropriate time.

5. The output of a cache

5.1 Definitions

Under the demand-driven caching operation (1.7), the output of the cache is the sequence of requests that incur a miss, i.e., when the incoming request cannot find the desired document in the cache. More precisely, a miss occurs at time t if R_t is *not* in S_t. Thus, we define recursively the time indices $\{\nu_k, \ k = 0, 1, \ldots\}$ by

$$\nu_0 = 0; \quad \nu_{k+1} := \nu_k + \mu_{k+1}, \quad k = 0, 1, \ldots$$

with

$$\mu_{k+1} := \inf\{\ell = 1, 2, \ldots : \ R_{\nu_k+\ell} \notin S_{\nu_k+\ell}\}$$

where we use the convention $\mu_{k+1} = \infty$ if either $\nu_k = \infty$ or if ν_k is finite but the set of indices entering the definition of μ_{k+1} is empty. With δ denoting an element *not* in $\mathcal{N}$, we define the output process $\{R_k^\star, \ k = 1, 2, \ldots\}$ simply as

$$R_k^\star := \begin{cases} R_{\nu_k} & \text{if } \nu_k < \infty \\ \delta & \text{if } \nu_k = \infty \end{cases}$$

for each $k = 1, 2, \ldots$. The requests $\{R_k^\star, k = 1, 2, \ldots\}$ are those requests among $\{R_t, t = 0, 1, \ldots\}$ which incur a miss and which get forwarded to the server (or to the higher level cache in a hierarchical caching system).

The statistics of the output stream $\{R^\star_k,\ k = 1, 2, \ldots\}$ are determined by the statistics of the input stream $\{R_t, t = 0, 1, \ldots\}$ and by the cache replacement policy π in use. We are interested in evaluating the popularity pmf $\boldsymbol{p}^\star_\pi = (p^\star_\pi(1), \ldots, p^\star_\pi(N))$ defined by

$$p^\star_\pi(i) := \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[R^\star_k = i\right] \quad a.s. \tag{1.12}$$

for each $i = 1, 2, \ldots, N$, whenever these limits exist.

5.2 Finding $\boldsymbol{p}^\star_\pi$

The remainder of this section is devoted to the existence and form of the limits (1.12).

Theorem 1.2 *Consider an eviction policy π such that the limits (1.9) exist under the IRM with popularity pmf $\boldsymbol{p}$. For each $i = 1, \ldots, N$, the limit (1.12) exists and is given by*

$$p^\star_\pi(i) \quad = \quad \frac{p(i)m_\pi(i;\boldsymbol{p})}{\sum_{j=1}^{N} p(j)m_\pi(j;\boldsymbol{p})} \tag{1.13}$$

where we have set

$$m_\pi(i;\boldsymbol{p}) := \sum_{s\in\Lambda^\star_i(M;\mathcal{N})} Q_\pi(s;\boldsymbol{p}). \tag{1.14}$$

A proof of Theorem 1.2 is given in Vanichpun and Makowski (2004b). Note that the existence of the limits (1.9) implies

$$\begin{aligned} m_\pi(i;\boldsymbol{p}) \quad &= \quad \sum_{s\in\Lambda^\star_i(M;\mathcal{N})} \left(\lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[S_\tau = s\right] \right) \\ &= \quad \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \sum_{s\in\Lambda^\star_i(M;\mathcal{N})} \mathbf{1}\left[S_\tau = s\right] \\ &= \quad \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[i \notin S_\tau\right] \quad a.s. \end{aligned} \tag{1.15}$$

for each $i = 1, \ldots, N$, and $m_\pi(i;\boldsymbol{p})$ thus represents the fraction of times that document i will not be in the cache. This quantity is determined by the popularity pmf $\boldsymbol{p}$ of the input to the cache and by the eviction policy π in use.

Inspection of (1.10) and (1.14) reveals that

$$\sum_{i=1}^{N} p(i) m_\pi(i; \boldsymbol{p}) = M_\pi(\boldsymbol{p})$$

and this leads via (1.13) to a simple connection between the miss rate of an eviction policy and the pmf of its output in the form

$$p^\star_\pi(i) = \frac{p(i) m_\pi(i; \boldsymbol{p})}{M_\pi(\boldsymbol{p})}, \quad i = 1, \ldots, N. \tag{1.16}$$

Thus, $p^\star_\pi(i)$ can be viewed as the ratio between the miss rate of the cache when the requested document is i and the overall miss rate of the cache.

6. Majorization – A primer

The concept of *majorization* provides a powerful tool to formalize statements concerning the relative skewness in the components of two vectors, viz., the components $(x_1, \ldots, x_N)$ of the vector $\boldsymbol{x}$ are "more spread out" or "more balanced" than the components $(y_1, \ldots, y_N)$ of the vector $\boldsymbol{y}$: For vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^N$, we say that $\boldsymbol{x}$ is *majorized* by $\boldsymbol{y}$, and write $\boldsymbol{x} \prec \boldsymbol{y}$, whenever the conditions

$$\sum_{i=1}^{n} x_{[i]} \leq \sum_{i=1}^{n} y_{[i]}, \quad n = 1, 2, \ldots, N-1 \tag{1.17}$$

and

$$\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i \tag{1.18}$$

hold with $x_{[1]} \geq x_{[2]} \geq \ldots \geq x_{[N]}$ and $y_{[1]} \geq y_{[2]} \geq \ldots \geq y_{[N]}$ denoting the components of $\boldsymbol{x}$ and $\boldsymbol{y}$ arranged in decreasing order, respectively.

We begin with a sufficient condition for majorization which is extracted from the discussion in Marshall and Olkin (1979), B.1, p. 129.

Proposition 1.3 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that (1.18) holds. Whenever, $x_1 \geq x_2 \geq \ldots \geq x_N$, if there exists some $k = 1, \ldots, N-1$ such that $x_i \leq y_i$, $i = 1, \ldots, k$, and $x_i \geq y_i$, $i = k+1, \ldots, N$, then the comparison $\boldsymbol{x} \prec \boldsymbol{y}$ holds.*

The following sufficient condition for majorization will be useful in the sequel; it was already announced in Marshall and Olkin (1979), B.1.b, p. 129, without proof.

Theorem 1.4 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that (1.18) holds. Whenever $x_1 \geq x_2 \geq \ldots \geq x_N > 0$, and the ratios $\frac{y_i}{x_i}$, $i = 1, \ldots, N$, are decreasing in i, we have the comparison $\boldsymbol{x} \prec \boldsymbol{y}$.*

With any element of $\mathbb{R}^N$ such that $\sum_{i=1}^N x_i \neq 0$, we associate the *normalized* vector $\bar{\boldsymbol{x}}$ as the element of $\mathbb{R}^N$ defined by

$$\bar{\boldsymbol{x}} := \big(\sum_{i=1}^N x_i\big)^{-1}(x_1, \ldots, x_N).$$

With this notation we can present a useful corollary to Theorem 1.4.

Corollary 1.5 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that $\sum_{i=1}^N y_i > 0$. Whenever $x_1 \geq x_2 \geq \ldots \geq x_N > 0$, and the ratios $\frac{y_i}{x_i}$, $i = 1, \ldots, N$, are decreasing in i, we have the comparison $\bar{\boldsymbol{x}} \prec \bar{\boldsymbol{y}}$.*

The following reformulation of Corollary 1.5 is used in the sequel.

Lemma 1.6 *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be distinct elements of $\mathbb{R}^N$ such that $x_i > 0$, $i = 1, \ldots, N$ and $\sum_{i=1}^N y_i > 0$. If $\frac{y_i}{x_i} \geq \frac{y_j}{x_j}$ whenever $x_i \geq x_j$ for distinct $i, j = 1, \ldots, N$, then the comparison $\bar{\boldsymbol{x}} \prec \bar{\boldsymbol{y}}$ holds.*

7. Schur-convexity

Key to the power of majorization is the companion notion of monotonicity associated with it: An $\mathbb{R}$-valued function φ defined on a set A of $\mathbb{R}^N$ is said to be Schur-convex (resp. Schur-concave) on A if

$$\varphi(\boldsymbol{x}) \leq \varphi(\boldsymbol{y}) \quad (\text{resp. } \varphi(\boldsymbol{x}) \geq \varphi(\boldsymbol{y}))$$

whenever $\boldsymbol{x}$ and $\boldsymbol{y}$ are elements in A satisfying $\boldsymbol{x} \prec \boldsymbol{y}$. In other words, Schur-convexity (resp. Schur-concavity) corresponds to monotone increasingness (resp. decreasingness) for majorization (viewed as a preorder on subsets of $\mathbb{R}^N$).

Let σ denote a permutation of $\{1, \ldots, N\}$. With any element $\boldsymbol{x}$ in $\mathbb{R}^N$, we associate the *permuted* vector $\sigma(\boldsymbol{x})$ in $\mathbb{R}^N$ through the relation

$$\sigma(\boldsymbol{x}) = (x_{\sigma(1)}, \ldots, x_{\sigma(N)}).$$

Let $\{\sigma_i,\ i = 1, \ldots, N!\}$ be a given enumeration of all the $N!$ permutations of $\{1, \ldots, N\}$; this enumeration is held fixed throughout the chapter. A subset A of $\mathbb{R}^N$ is said to be *symmetric* if for any $\boldsymbol{x}$ in A, the element $\sigma_i(\boldsymbol{x})$ also belongs to A for *each* $i = 1, \ldots, N!$. Moreover, for any subset A of $\mathbb{R}^N$, a mapping $\varphi : A \to \mathbb{R}$ is said to be *symmetric* if A is symmetric and for any $\boldsymbol{x}$ in A, we have $\varphi(\sigma_i(\boldsymbol{x})) = \varphi(\boldsymbol{x})$ for *each* $i = 1, \ldots, N!$. If the mapping $\varphi : A \to \mathbb{R}$ is Schur-convex (resp. Schur-concave) with symmetric A, then φ is necessarily symmetric since $\sigma_i(\boldsymbol{x}) \prec \boldsymbol{x} \prec \sigma_i(\boldsymbol{x})$ implies $\varphi(\sigma_i(\boldsymbol{x})) = \varphi(\boldsymbol{x})$ for each $i = 1, \ldots, N!$.

In the following, we have collected some useful technical results concerning Schur concavity. As in Marshall and Olkin (1979), p. 78, for each $M = 1, \ldots, N$, the *elementary symmetric* function $E_{M,N} : \mathbb{R}^N \to \mathbb{R}$ is defined by

$$E_{M,N}(\boldsymbol{x}) := \sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(M;\mathcal{N})} x_{i_1} \cdots x_{i_M}, \quad \boldsymbol{x} \in \mathbb{R}^N. \tag{1.19}$$

By convention we write $E_{0,N}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x}$ in $\mathbb{R}^N$. It is well known (Marshall and Olkin (1979), Prop. F.1., p. 78) that the function $E_{M,N}$ is Schur-concave on $\mathbb{R}^N_+$ for each $M = 0, 1, \ldots, N$.

The following result is due to Schur (see Marshall and Olkin (1979), F.3, p. 80) and will be key to a number of proofs.

Proposition 1.7 *For each $M = 1, \ldots, N$, the mapping $\Phi_{M,N} : \mathbb{R}^N_+ \to \mathbb{R}$ given by*[5]

$$\Phi_{M,N}(\boldsymbol{x}) := \frac{E_{M,N}(\boldsymbol{x})}{E_{M-1,N}(\boldsymbol{x})}, \quad \boldsymbol{x} \in \mathbb{R}^N_+$$

is increasing,[6] *symmetric and concave, thus increasing and Schur-concave on $\mathbb{R}^N_+$.*

With vectors $\boldsymbol{t}$ and $\boldsymbol{x}$ in $\mathbb{R}^N$, we associate the element $\boldsymbol{t} \cdot \boldsymbol{x}$ of $\mathbb{R}^N$ defined by $\boldsymbol{t} \cdot \boldsymbol{x} := (t_1 x_1, \ldots, t_N x_N)$. With this notation we can state

Proposition 1.8 *Assume the mapping $\psi : \mathbb{R}^N_+ \to \mathbb{R}$ to be concave and the mapping $h : \mathbb{R}^{N!} \to \mathbb{R}$ to be increasing, symmetric and concave. For any non-zero vector $\boldsymbol{t}$ in $\mathbb{R}^N$, the mapping $\psi_{\boldsymbol{t}} : \mathbb{R}^N_+ \to \mathbb{R}$ defined by*

$$\psi_{\boldsymbol{t}}(\boldsymbol{x}) := h(\psi(\boldsymbol{t} \cdot \sigma_1(\boldsymbol{x})), \ldots, \psi(\boldsymbol{t} \cdot \sigma_{N!}(\boldsymbol{x}))), \quad \boldsymbol{x} \in \mathbb{R}^N_+$$

is symmetric and concave, thus Schur-concave on $\mathbb{R}^N_+$.

8. Comparing input and output

Recall that we have in mind to compare the strength of locality of reference in two streams of requests through a majorization ordering of

[5] For $\boldsymbol{x}$ in $\mathbb{R}^N_+$ such that $E_{M-1,N}(\boldsymbol{x}) = 0$, then $E_{M,N}(\boldsymbol{x}) = 0$ and we set $\Phi_{M,N}(\boldsymbol{x}) = 0$ by continuity.

[6] Here, increasing means increasing in each argument.

their popularity pmfs. The next result constitutes a first step in the process of comparing input and output popularity pmfs.

THEOREM 1.9 *Consider an eviction policy π such that the limits (1.9) exist under the IRM with popularity pmf $\boldsymbol{p}$. If $m_\pi(i;\boldsymbol{p}) \geq m_\pi(j;\boldsymbol{p})$ whenever $p(i)m_\pi(i;\boldsymbol{p}) \leq p(j)m_\pi(j;\boldsymbol{p})$ for distinct $i, j = 1, \ldots, N$, then it holds that $\boldsymbol{p}^\star_\pi \prec \boldsymbol{p}$ provided $m_\pi(i;\boldsymbol{p}) > 0$ for each $i = 1, \ldots, N$.*

Proof. This claim is a simple consequence of Lemma 1.6: We take $\boldsymbol{y} = \boldsymbol{p}$ and $\boldsymbol{x}$ given by $x_i = p(i)m_\pi(i;\boldsymbol{p})$, $i = 1, \ldots, N$. Thus, we have $\bar{\boldsymbol{x}} = \boldsymbol{p}^\star_\pi$ while $\bar{\boldsymbol{y}} = \boldsymbol{p}$, and the requisite monotonicity assumptions hold. ■

The assumptions of Theorem 1.9 ensure that $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ and $p(j) \leq p(i)$ occur simultaneously for distinct $i, j = 1, \ldots, N$. This leads to defining a caching algorithm π as *good* if for every admissible pmf $\boldsymbol{p}$, we have $m_\pi(i;\boldsymbol{p}) \leq m_\pi(j;\boldsymbol{p})$ whenever $p(j) \leq p(i)$ for distinct $i, j = 1, \ldots, N$. Thus, a caching policy which satisfies the assumptions of Theorem 1.9 is necessarily a good policy. However, as we shall see in the case of the LRU policy, this by itself is not sufficient to ensure that the output popularity pmf is more balanced than the input popularity pmf.

Repeatedly we shall encounter output pmfs which assume the generic form used in Theorem 1.10 below.

THEOREM 1.10 *Let $\boldsymbol{p}$ be an admissible pmf on $\mathcal{N}$, and for each $i = 1, \ldots, N$, define an $(N-1)$-dimensional vector*

$$\boldsymbol{p}^{(i)} := (p(1), \ldots, p(i-1), p(i+1), \ldots, p(N)).$$

For each $M = 1, 2, \ldots, N-1$, the pmf $\boldsymbol{p}^\star_M$ on $\mathcal{N}$ defined by

$$p^\star_M(i) = \frac{p(i)E_{M,N-1}(\boldsymbol{p}^{(i)})}{\sum_{j=1}^N p(j)E_{M,N-1}(\boldsymbol{p}^{(j)})} \quad i = 1, \ldots, N \tag{1.20}$$

satisfies the comparison $\boldsymbol{p}^\star_M \prec \boldsymbol{p}$.

A proof of this theorem builds on Lemma 1.6 and is given in Vanichpun and Makowski (2004b).

9. Random on-demand replacement

We now introduce a large class of demand-driven eviction policies called *Random On-demand Replacement Algorithms* (RORA). This class of policies generalizes many well-known caching policies, e.g., the random and FIFO policies, as well as the optimal policy A_0. Moreover, the Partially Preloaded Random Replacement Algorithms proposed by Gelenbe (1973) form a subclass of RORAs.

9.1 Defining RORAs

A RORA policy follows the demand-driven caching rule (1.7) (under the customary assumption that the cache is initially full) and is characterized by an *eviction/insertion* pmf $\boldsymbol{r}$ which we organize as the $M \times M$ matrix $\boldsymbol{r} = (r_{k\ell})$, i.e., for each $k, \ell = 1, \ldots, M$, we have $r_{k\ell} \geq 0$ and $\sum_{k=1}^{M} \sum_{\ell=1}^{M} r_{k\ell} = 1$. The RORA associated with the pmf matrix $\boldsymbol{r}$ is denoted by RORA($\boldsymbol{r}$).

We select the cache state Ω_t at time t to be an element $(i_1, \ldots, i_M)$ of $\Lambda(M; \mathcal{N})$ with the understanding that document i_k, $k = 1, \ldots, M$, is in cache position k at time t. RORA($\boldsymbol{r}$) implements the following eviction rule: Introduce a sequence of i.i.d. rvs $\{(X_t, Y_t),\ t = 0, 1, \ldots\}$ taking values in $\{1, \ldots, M\} \times \{1 \ldots, M\}$ with common pmf $\boldsymbol{r}$, i.e., for each $t = 0, 1, \ldots$, we have

$$\mathbf{P}\left[(X_t, Y_t) = (k, \ell)\right] = r_{k\ell}, \quad k, \ell = 1, \ldots, M.$$

The sequences of rvs $\{(X_t, Y_t),\ t = 0, 1, \ldots\}$ and $\{R_t,\ t = 0, 1, \ldots\}$ are assumed mutually independent. The document U_t to be evicted at time t is given by

$$U_t = \mathbf{1}\left[R_t \notin S_t\right] i_{X_t}.$$

We have $U_t = 0$ whenever $R_t \in S_t$, in which case no replacement occurs and the cache state remains unchanged, i.e., $\Omega_{t+1} = \Omega_t$.

Next, if $R_t \notin S_t$ and $(X_t, Y_t) = (k, \ell)$, then $U_t = i_k$ (the document at position k is evicted) and the new document is inserted in the cache at position ℓ. If $k < \ell$, the documents $i_{k+1}, \ldots, i_\ell$ are shifted down to position $k, k+1 \ldots, \ell - 1$ (in that order) while if $k > \ell$, the documents $i_\ell, \ldots, i_{k-1}$ are shifted up to position $\ell + 1, \ldots, k$ (in that order). When $k = \ell$, the new document simply replaces the evicted document at position k.

A document initially at position i in the cache will *never* be replaced if

$$r_{k\ell} = 0 \quad \text{for } k \leq i \leq \ell \quad \text{and} \quad \ell \leq i \leq k. \tag{1.21}$$

If we use row i and column i to partition the matrix $\boldsymbol{r}$ into four blocks, then condition (1.21) expresses the fact that the entries in the northwest and southeast corners *all* vanish (including row i and column i). Let $\Sigma_{\boldsymbol{r}}$ denote the set of *cache positions* with the property that any document initially put there will never be evicted during the operation of the cache, i.e.,

$$\Sigma_{\boldsymbol{r}} := \{i = 1, \ldots, M : \text{ Eqn. (1.21) holds at } i\}. \tag{1.22}$$

Under the IRM with popularity pmf $\boldsymbol{p}$, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form a Markov chain on the state space $\Lambda(M; \mathcal{N})$. The ergodic properties of this chain are determined by whether the set $\Sigma_{\boldsymbol{r}}$ is empty or not. This is discussed in Lemmas 1.11 and 1.12 in the next two sections; they are established in Vanichpun (2004).

Throughout this discussion we always assume that the cache size M and the number of cacheable documents N satisfy $M + 1 < N$. We do so in order to avoid technical cases of limited interest. Indeed, the results here are still valid for the case $N = M + 1$, but require slightly different arguments. We refer the interested reader to Vanichpun (2004).

9.2 Case 1

The set $\Sigma_{\boldsymbol{r}}$ is *empty*, so that *every* document in cache is eventually replaced, i.e., for each $i = 1, \ldots, M$, there exists a pair k, ℓ (possibly depending on i) with either $1 \leq k \leq i \leq \ell \leq M$ or $1 \leq \ell \leq i \leq k \leq M$ such that $r_{k\ell} > 0$. Here are some well-known policies which fall in this case: The random policy corresponds to RORA($\boldsymbol{r}$) with $\boldsymbol{r}$ given by $r_{kk} = \frac{1}{M}$ for each $k = 1, \ldots, M$. The FIFO policy also belongs to RORA with two possibilities for $\boldsymbol{r}$, namely $r_{1M} = 1$ or $r_{M1} = 1$. The first (resp. second) choice corresponds to the cache state $(i_1, \ldots, i_M)$ being loaded from left to right with documents ordered from the oldest to the most recent (resp. from the most recent to the oldest).

In this case, the Markov chain $\{\Omega_t, t = 0, 1, \ldots\}$ is ergodic on the state space $\Lambda(M; \mathcal{N})$; its stationary distribution exists and is given in the following lemma.

Lemma 1.11 *Assume the input to be an IRM with popularity pmf $\boldsymbol{p}$. For RORA($\boldsymbol{r}$) with $\Sigma_{\boldsymbol{r}}$ empty, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ is an ergodic Markov chain on the state space $\Lambda(M; \mathcal{N})$ with stationary pmf on $\Lambda(M; \mathcal{N})$ given by*

$$\begin{aligned}\pi_{\boldsymbol{r}}(s; \boldsymbol{p}) &= \lim_{t \to \infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s. \\ &= C(\boldsymbol{p})^{-1} p(i_1) p(i_2) \cdots p(i_M) \qquad (1.23)\end{aligned}$$

for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$ with normalizing constant

$$C(\boldsymbol{p}) := \sum_{(i_1, \ldots, i_M) \in \Lambda(M; \mathcal{N})} p(i_1) p(i_2) \cdots p(i_M). \tag{1.24}$$

The stationary pmf is the same for *all* RORAs in Case 1.

9.3 Case 2

The set $\Sigma_{\boldsymbol{r}}$ is *not* empty, and some documents, once put in cache, will never be replaced during the operation of the cache, i.e., if $\Omega_0 = (i_1, \ldots, i_M)$, then for all $t = 1, 2, \ldots$, with $\Omega_t = (j_1, \ldots, j_M)$, we have

$$j_\ell = i_\ell, \quad \ell \in \Sigma_{\boldsymbol{r}}. \tag{1.25}$$

Here are some examples of RORA policies in that category: As pointed out in Section 1.3.3, any permutation σ of $\{1, \ldots, N\}$ induces an eviction policy A_σ which evicts the "smallest" document in cache with documents $\sigma(1), \sigma(2), \ldots, \sigma(N)$ "ranked" in decreasing order. The documents $\sigma(1), \ldots, \sigma(M-1)$, once loaded in the cache, will remain there. This behavior can be recovered through the RORA($\boldsymbol{r}$) policy with matrix $\boldsymbol{r}$ of the form $r_{kk} = 1$ for some $k = 1, \ldots, M$, in which case $\Sigma_{\boldsymbol{r}}$ has $M-1$ elements, namely $\{1, \ldots, k-1, k+1, \ldots, M\}$. If the documents $\sigma(1), \ldots, \sigma(M-1)$ are initially put in cache (i.e., preloaded) at the other positions $\ell \neq k$ in $\Sigma_{\boldsymbol{r}}$, this RORA($\boldsymbol{r}$) policy will behave like the policy A_σ in its *steady state* regime. The steady state behavior of the cache under the policy A_0 introduced in Section 1.3.3, is that of the RORA($\boldsymbol{r}$) above, this time, the preloaded documents being the $M-1$ *most popular* documents.

To describe the long-run behavior of the cache states $\{\Omega_t, t = 0, 1, \ldots\}$, we go back to (1.25). First, with initial cache state $s_0 = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$, let $\Sigma_{\boldsymbol{r}}(s_0)$ denote the set of initial documents with positions in $\Sigma_{\boldsymbol{r}}$, i.e.,

$$\Sigma_{\boldsymbol{r}}(s_0) := \{i_\ell : \ \ell \in \Sigma_{\boldsymbol{r}}\}. \tag{1.26}$$

Next, we introduce the component

$$\Lambda(\boldsymbol{r}, s_0) := \{(j_1, \ldots, j_M) \in \Lambda(M; \mathcal{N}) : \ j_\ell = i_\ell, \ \ell \in \Sigma_{\boldsymbol{r}}\}. \tag{1.27}$$

In view of (1.25), once the cache state is in $\Lambda(\boldsymbol{r}, s_0)$, it remains there forever. In fact *all* the states in the component $\Lambda(\boldsymbol{r}, s_0)$ communicate with each other, and this set of states is closed under the motion of the Markov chain. There are $\binom{N-m}{M-m}(M-m)!$ elements in $\Lambda(\boldsymbol{r}, s_0)$ and there are $\binom{N}{m}m!$ *distinct* components which form a partition of $\Lambda(M; \mathcal{N})$. As a result, when restricted to $\Lambda(\boldsymbol{r}, s_0)$, this Markov chain is irreducible and aperiodic, and its ergodic behavior can be characterized as follows:

Lemma 1.12 *Assume the input to be an IRM with popularity pmf* $\boldsymbol{p}$. *For RORA*($\boldsymbol{r}$) *with* $|\Sigma_{\boldsymbol{r}}| = m$ *for some* $m = 1, \ldots, M-1$, *and initial* *cache state s_0, the cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form an ergodic Markov chain on the component $\Lambda(\boldsymbol{r}, s_0)$. In particular the limit*

$$\pi_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s.$$

always exists for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M;\mathcal{N})$ with

$$\pi_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) = C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \tag{1.28}$$

for every s in $\Lambda(\boldsymbol{r}, s_0)$, and $\pi_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) = 0$ otherwise, with normalizing constant

$$C_{\boldsymbol{r}}(\boldsymbol{p}, s_0) := \sum_{(i_1,\ldots,i_M)\in\Lambda(\boldsymbol{r},s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell). \tag{1.29}$$

10. The miss rate under RORAs

10.1 Case 1

Fix $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M;\mathcal{N})$, and let $\Lambda(s|M;\mathcal{N})$ denote the subset of $\Lambda(M;\mathcal{N})$ defined by

$$\Lambda(s|M;\mathcal{N}) := \{(j_1, \ldots, j_M) \in \Lambda(M;\mathcal{N}) : \ \{j_1, \ldots, j_M\} = \{i_1, \ldots, i_M\}\}.$$

By Lemma 1.11, the limit (1.9) exists and is given by

$$\begin{aligned} Q_{\boldsymbol{r}}(s;\boldsymbol{p}) &= \sum_{(j_1,\ldots,j_M)\in\Lambda(s|M;\mathcal{N})} C(\boldsymbol{p})^{-1} p(j_1)p(j_2)\cdots p(j_M) \\ &= C(\boldsymbol{p})^{-1} M! \cdot p(i_1)p(i_2)\cdots p(i_M) \end{aligned} \tag{1.30}$$

with normalizing constant $C(\boldsymbol{p})$ given by (1.24). The last equality at (1.30) follows from the fact that there are $M!$ elements in $\Lambda(s|M;\mathcal{N})$.

Using (1.30) in conjunction with Theorem 1.1, we readily conclude that under the RORA($\boldsymbol{r}$) policy of Case 1 the miss rate (1.8) exists as a constant which is independent of the initial cache state s_0. To acknowledge this fact, we simply denote this limiting constant by $M_{\boldsymbol{r}}(\boldsymbol{p})$. Specializing (1.11) leads to

$$\begin{aligned} M_{\boldsymbol{r}}(\boldsymbol{p}) &= C(\boldsymbol{p})^{-1} M! \sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M) \sum_{i\notin\{i_1,\ldots,i_M\}} p(i) \\ &= C(\boldsymbol{p})^{-1} (M+1)! \sum_{\{i_1,\ldots,i_{M+1}\}\in\Lambda^\star(M+1;\mathcal{N})} p(i_1)\cdots p(i_{M+1}) \\ &= C(\boldsymbol{p})^{-1} (M+1)! \cdot E_{M+1,N}(\boldsymbol{p}) \end{aligned} \tag{1.31}$$

while the normalizing constant $C(\boldsymbol{p})$ in (1.24) can be simplified as

$$\begin{aligned}\sum_{(i_1,\ldots,i_M)\in\Lambda(M;\mathcal{N})} p(i_1)\cdots p(i_M) &= M! \sum_{\{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N})} p(i_1)\cdots p(i_M) \\ &= M!\cdot E_{M,N}(\boldsymbol{p}). \qquad (1.32)\end{aligned}$$

Combining (1.31) and (1.32) we finally get

$$M_{\boldsymbol{r}}(\boldsymbol{p}) = (M+1)\cdot\frac{E_{M+1,N}(\boldsymbol{p})}{E_{M,N}(\boldsymbol{p})} = (M+1)\Phi_{M+1,N}(\boldsymbol{p}) \qquad (1.33)$$

and a straightforward application of Proposition 1.7 yields

THEOREM 1.13 *Under the RORA($\boldsymbol{r}$) policy with $\Sigma_{\boldsymbol{r}}$ empty, for admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that $M_{\boldsymbol{r}}(\boldsymbol{q}) \leq M_{\boldsymbol{r}}(\boldsymbol{p})$ whenever $\boldsymbol{p} \prec \boldsymbol{q}$.*

10.2 Case 2

Consider now the RORA($\boldsymbol{r}$) policy when the set $\Sigma_{\boldsymbol{r}}$ is *not* empty, say with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1,\ldots,M-1$, and let the cache be initially in state s_0 in $\Lambda(M;\mathcal{N})$. By Lemma 1.12, for each $s = \{i_1,\ldots,i_M\}$ in $\Lambda^\star(M;\mathcal{N})$ the limit (1.9) exists and is given by

$$Q_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) = \sum_{s'\in\Lambda(s|\boldsymbol{r},s_0)} \pi_{\boldsymbol{r},s_0}(s';\boldsymbol{p}) \qquad (1.34)$$

where $\Lambda(s|\boldsymbol{r},s_0)$ denotes the subset of $\Lambda(\boldsymbol{r},s_0)$ defined by

$$\Lambda(s|\boldsymbol{r},s_0) := \{(j_1,\ldots,j_M)\in\Lambda(\boldsymbol{r},s_0):\ \{j_1,\ldots,j_M\} = \{i_1,\ldots,i_M\}\}.$$

The set $\Lambda(s|\boldsymbol{r},s_0)$ is *non*-empty if and only if

$$\Sigma_{\boldsymbol{r}}(s_0) \subseteq \{i_1,\ldots,i_M\} \qquad (1.35)$$

so that $Q_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) = 0$ whenever this inclusion (1.35) does not hold. With this in mind we define

$$\Lambda^\star(\boldsymbol{r},s_0) := \{s = \{i_1,\ldots,i_M\}\in\Lambda^\star(M;\mathcal{N}):\ \text{Eqn. (1.35) holds at } s\}.$$

Going back to (1.28) and (1.29), we now conclude that for each $s = \{i_1,\ldots,i_M\}$ in $\Lambda^\star(\boldsymbol{r},s_0)$, it holds

$$\begin{aligned}Q_{\boldsymbol{r},s_0}(s;\boldsymbol{p}) &= \sum_{(j_1,\ldots,j_M)\in\Lambda(s|\boldsymbol{r},s_0)} C_{\boldsymbol{r}}(\boldsymbol{p},s_0)^{-1}\prod_{j_\ell\notin\Sigma_{\boldsymbol{r}}(s_0)} p(j_\ell) \\ &= C_{\boldsymbol{r}}(\boldsymbol{p},s_0)^{-1}(M-m)!\cdot\prod_{i_\ell\notin\Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \qquad (1.36)\end{aligned}$$

where in the last equality we combine the set equality $\{j_1, \ldots, j_M\} = \{i_1, \ldots, i_M\}$ with (1.35), and then made use of the identity $|\Lambda(s|\boldsymbol{r}, s_0)| = (M-m)!$.

Now, using (1.36) in conjunction with Theorem 1.1, we see that under the RORA($\boldsymbol{r}$) policy of Case 2 the miss rate (1.8) exists as a constant which *depends* on the initial cache state s_0. We record this fact in the notation by denoting this limiting constant by $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0)$. As in Case 1, specializing (1.11) leads to

$$\begin{aligned} & M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) \\ = \; & C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M-m)! \sum_{\{i_1,\ldots,i_M\} \in \Lambda^\star(\boldsymbol{r}, s_0)} \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \sum_{i \notin \{i_1,\ldots,i_M\}} p(i) \\ = \; & C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M-m+1)! \cdot E_{M-m+1,N}(\boldsymbol{t} \cdot \boldsymbol{p}) \qquad (1.37) \end{aligned}$$

where the element $\boldsymbol{t}$ in $\mathbb{R}^N$ is specified by $t_i = 0$ for document i in $\Sigma_{\boldsymbol{r}}(s_0)$ and $t_i = 1$ otherwise. Moreover, by the same arguments as in Case 1, we can simplify the normalizing constant $C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)$ as

$$C_{\boldsymbol{r}}(\boldsymbol{p}, s_0) = (M-m)! \cdot E_{M-m,N}(\boldsymbol{t} \cdot \boldsymbol{p}). \qquad (1.38)$$

It then follows from (1.37) and (1.38) that

$$\begin{aligned} M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) &= (M-m+1) \cdot \frac{E_{M-m+1,N}(\boldsymbol{t} \cdot \boldsymbol{p})}{E_{M-m,N}(\boldsymbol{t} \cdot \boldsymbol{p})} \\ &= (M-m+1)\Phi_{M-m+1,N}(\boldsymbol{t} \cdot \boldsymbol{p}). \qquad (1.39) \end{aligned}$$

Clearly, the documents in $\Sigma_{\boldsymbol{r}}(s_0)$ do not contribute to the miss rate since they never generate a miss once loaded in cache – This is *regardless* of the order in which they appear in the cache state s_0. This intuitively obvious fact is in agreement with the expression (1.39) from which we see that for any two initial cache states s_0 and s'_0 in $\Lambda(M; \mathcal{N})$ with $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma_{\boldsymbol{r}}(s'_0)$, we have the equality $M_{\boldsymbol{r}}(\boldsymbol{p}; s_0) = M_{\boldsymbol{r}}(\boldsymbol{p}; s'_0)$. As a result, we shall find it appropriate to denote this common value by $M_{\boldsymbol{r},\Sigma_{\boldsymbol{r}}(s_0)}(\boldsymbol{p})$.

For any pmf $\boldsymbol{p}$ on $\mathcal{N}$, let $\Sigma^\star(\boldsymbol{p})$ denote the set of the m most popular documents according to the pmf $\boldsymbol{p}$. Equipped with the expression (1.39), we are now ready to establish the result for RORA policies in Case 2.

THEOREM 1.14 *Under the RORA($\boldsymbol{r}$) policy with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M-1$, for admissible pmfs $\boldsymbol{p}$ and $\boldsymbol{q}$ on $\mathcal{N}$, it holds that*

$$M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{q})}(\boldsymbol{q}) \leq M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{p})}(\boldsymbol{p}) \qquad (1.40)$$

whenever $\boldsymbol{p} \prec \boldsymbol{q}$.

Proof. The desired result will be established if we can show that the miss rate function $\boldsymbol{p} \to M_{\boldsymbol{r},\Sigma_{\boldsymbol{r}}(s_0)}(\boldsymbol{p})$ as given in (1.39) is Schur-concave whenever s_0 is selected so that $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma^\star(\boldsymbol{p})$.

As we can always relabel the documents, there is no loss of generality in assuming $p(1) \geq p(2) \geq \ldots \geq p(N)$, whence $\Sigma^\star(\boldsymbol{p}) = \{1, \ldots, m\}$ and the element $\boldsymbol{t}$ in (1.39) can be specified as $t_1 = \ldots = t_m = 0$ and $t_{m+1} = \ldots = t_N = 1$. By Proposition 1.7, the mapping $\Phi_{M-m+1,N}$ is increasing and Schur-concave on $\mathbb{R}_+^N$, and by virtue of the defining property of $\Sigma^\star(\boldsymbol{p})$, we have

$$M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{p})}(\boldsymbol{p}) \quad = \quad (M-m+1) \cdot \min_{i=1,\ldots,N!} \Phi_{M-m+1,N}(\boldsymbol{t} \cdot \sigma_i(\boldsymbol{p})).$$

The mapping $h : \mathbb{R}^{N!} \to \mathbb{R} : \boldsymbol{y} \to \min(y_1, \ldots, y_{N!})$ is clearly increasing, symmetric and concave, while the mapping $\Phi_{M-m+1,N}$ is concave on $\mathbb{R}_+^N$ by Proposition 1.7. Combining these facts with the expression for $M_{\boldsymbol{r},\Sigma^\star(\boldsymbol{p})}(\boldsymbol{p})$ obtained above, we conclude by Proposition 1.8 to the Schur-concavity (in the pmf vector) of the miss rate functional (1.39) under the RORA policy when $\Sigma_{\boldsymbol{r}}(s_0) = \Sigma^\star(\boldsymbol{p})$. ■

11. The output under RORAs

We now discuss the popularity pmf of the output generated under the RORA policies.

11.1 Case 1

As we invoke Theorem 1.2, we can make use of the expressions (1.30) into the relation (1.14). For each $i = 1, \ldots, N$, in the notation of Theorem 1.10, this yields

$$\begin{aligned} m_{\boldsymbol{r}}(i;\boldsymbol{p}) &= \sum_{s \in \Lambda_i^\star(M;\mathcal{N})} C(\boldsymbol{p})^{-1} M! \cdot p(i_1)p(i_2)\cdots p(i_M) \\ &= \frac{E_{M,N-1}(\boldsymbol{p}^{(i)})}{E_{M,N}(\boldsymbol{p})} \qquad (1.41) \end{aligned}$$

where the last equality follows from (1.32).

Reporting (1.41) back into (1.13), we conclude that the popularity pmf $\boldsymbol{p}_{\boldsymbol{r}}^\star$ of the output produced by RORA($\boldsymbol{r}$) policy in Case 1 is indeed of the form (1.20), and Theorem 1.10 gives us

THEOREM 1.15 *Under the RORA($\boldsymbol{r}$) policy in Case 1, it holds that* $\boldsymbol{p}_{\boldsymbol{r}}^\star \prec \boldsymbol{p}$.

When $M = 1$, any demand-driven policy π reduces to the policy that evicts the only document in cache if the requested document is not in cache. Specializing the results above. we find that the output pmf $\boldsymbol{p}^\star_\pi$ is given by

$$p^\star_\pi(i) = \frac{p(i)(1-p(i))}{\sum_{j=1}^N p(j)(1-p(j))}, \quad i = 1, \ldots, N \tag{1.42}$$

and Theorem 1.15 immediately leads to

Corollary 1.16 *With $M = 1$, under any demand-driven replacement policy π, the popularity pmf $\boldsymbol{p}^\star_\pi$ of the output is given by (1.42), and satisfies $\boldsymbol{p}^\star_\pi \prec \boldsymbol{p}$.*

11.2 Case 2

Assume $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M-1$, and let the cache be initially in state s_0. The pmf $\boldsymbol{\pi}$ on $\Sigma_{\boldsymbol{r}}(s_0)^c$ is defined as the *conditional* pmf induced by $\boldsymbol{p}$ on $\Sigma_{\boldsymbol{r}}(s_0)^c$; it is given by

$$\pi(i) = \frac{p(i)}{\sum_{j \in \Sigma_{\boldsymbol{r}}(s_0)^c} p(j)}, \quad i \in \Sigma_{\boldsymbol{r}}(s_0)^c. \tag{1.43}$$

For all i in $\Sigma_{\boldsymbol{r}}(s_0)$, it is clear that $m_{\boldsymbol{r},s_0}(i;\boldsymbol{p}) = 0$ while for document i *not* in $\Sigma_{\boldsymbol{r}}(s_0)$, with the expression for $Q_{\boldsymbol{r},s_0}(s;\boldsymbol{p})$ given in (1.36), we find

$$\begin{aligned} m_{\boldsymbol{r},s_0}(i;\boldsymbol{p}) &= \sum_{s \in \Lambda^\star(\boldsymbol{r},s_0):\ i \notin s} C_{\boldsymbol{r}}(\boldsymbol{p}, s_0)^{-1}(M-m)! \cdot \prod_{i_\ell \notin \Sigma_{\boldsymbol{r}}(s_0)} p(i_\ell) \\ &= \frac{E_{M-m,N}(\boldsymbol{t}^{(1)} \cdot \boldsymbol{p})}{E_{M-m,N}(\boldsymbol{t}^{(2)} \cdot \boldsymbol{p})} \\ &= \frac{E_{M-m,N-m-1}(\boldsymbol{\pi}^{(i)})}{E_{M-m,N-m}(\boldsymbol{\pi})} \end{aligned} \tag{1.44}$$

where the element $\boldsymbol{t}^{(1)}$ and $\boldsymbol{t}^{(2)}$ of $\mathbb{R}^N$ are specified by $t^{(1)}_j = t^{(2)}_j = 0$ for document j in $\Sigma_{\boldsymbol{r}}(s_0)$, $t^{(1)}_i = 0$, $t^{(2)}_i = 1$ and $t^{(1)}_j = t^{(2)}_j = 1$ whenever document $j \neq i$ is *not* in $\Sigma_{\boldsymbol{r}}(s_0)$. In the second equality we made use of the expression (1.38).

Combining (1.44) with (1.13), we immediately get

$$p^\star_{\boldsymbol{r},s_0}(i) = \begin{cases} 0 & \text{if } i \in \Sigma(s_0) \\ \frac{\pi(i) E_{M-m,N-m-1}(\boldsymbol{\pi}^{(i)})}{\sum_{j \in \Sigma(s_0)^c} \pi(j) E_{M-m,N-m-1}(\boldsymbol{\pi}^{(j)})} & \text{if } i \notin \Sigma(s_0). \end{cases} \tag{1.45}$$

Since $p^\star_{\boldsymbol{r},s_0}(i) = 0$ whenever i belongs to $\Sigma_{\boldsymbol{r}}(s_0)$, it is more natural to seek a comparison between $\boldsymbol{p}^\star_{\boldsymbol{r},s_0}$ and the conditional pmf $\boldsymbol{\pi}$.

Theorem 1.17 *Under the RORA($\boldsymbol{r}$) policy with $|\Sigma_{\boldsymbol{r}}| = m$ for some $m = 1, \ldots, M-1$, it holds that $\boldsymbol{p}^\star_{\boldsymbol{r},s_0} \prec \boldsymbol{\pi}$.*

Theorem 1.17 is essentially the same as Theorem 1.10. We immediately obtain the desired result upon identifying $\boldsymbol{\pi}$ and $\Sigma_{\boldsymbol{r}}(s_0)^c$ with $\boldsymbol{p}$ and $\mathcal{N}$ in Theorem 1.10, respectively.

12. Zipf-like pmfs

It has been observed in a number of studies that the popularity distribution of objects in request streams at Web caches is highly skewed. In Almeida et al. (1996), a good fit was provided by the *Zipf* distribution according to which the popularity of the i^{th} most popular object is inversely proportional to its rank, namely $1/i$.

In more recent studies by Breslau et al. (1999) and by Jin and Bestavros (2000a), "Zipf-like" distributions[7] were found more appropriate; see Breslau et al. (1999) (and references therein) for an excellent summary. Such distributions form a one-parameter family. In our set-up, we say that the popularity pmf $\boldsymbol{p}$ of the $\mathcal{N}$-valued rvs $\{R_t,\ t = 0, 1, \ldots\}$ is Zipf-like with parameter $\alpha \geq 0$ if

$$p(i) = \frac{i^{-\alpha}}{C_\alpha(N)}, \quad i = 1, \ldots, N \quad \text{with} \quad C_\alpha(N) := \sum_{i=1}^{N} i^{-\alpha}. \tag{1.46}$$

The pmf (1.46) will be denoted by $\boldsymbol{p}_\alpha$. It is always the case that $p_\alpha(1) \geq p_\alpha(2) \geq \ldots \geq p_\alpha(N)$. The case $\alpha = 1$ corresponds to the standard Zipf distribution and as studied by Breslau et al. (1999), the value of α was typically found to be in the range $0.64 - 0.83$.

Zipf-like pmfs are skewed towards the most popular objects. As $\alpha \to 0$, the Zipf-like pmf approaches the uniform distribution $\boldsymbol{u}$ while as $\alpha \to \infty$, it degenerates to the pmf $(1, 0, \ldots, 0)$. Extrapolating between these extreme cases, we expect the parameter α of Zipf-like pmfs (1.46) to measure the strength of skewness, with the larger α, the more skewed the pmf $\boldsymbol{p}_\alpha$. The next result shows that majorization indeed captures this fact, and so it is warranted to call α the *skewness parameter* of the Zipf-like pmf.

Lemma 1.18 *For $0 \leq \alpha < \beta$, it holds that $\boldsymbol{p}_\alpha \prec \boldsymbol{p}_\beta$.*

Lemma 1.18 can already be found in Marshall and Olkin (1979), B.2.b, p. 130, and is an easy by-product of Lemma 1.6. In the spirit of Lemma

[7] Such distributions are sometimes called generalized Zipf distributions.

1.18 and the aforementioned folk theorem (1.3), we expect the miss rate of the cache replacement policy to decrease as α increases. This has been shown to be the case using simulations in Gadde, Chase and Rabinovich (2001).

Zipf-like pmfs are used in the discussion of the LRU policy in the next sections.

13. The miss rate under the LRU policy

Under the IRM with admissible popularity pmf $\boldsymbol{p}$, it is known (Aven, Coffman and Kogan (1987), Thm. 9, p. 130 and Coffman and Denning (1973), Thm. 6.5, p. 272) that the LRU cache states $\{\Omega_t, t = 0, 1, \ldots\}$ form a stationary ergodic Markov chain over the finite state space $\Lambda(M; \mathcal{N})$ with stationary distribution given by

$$\begin{aligned}\pi_{\mathrm{LRU}}(s; \boldsymbol{p}) &= \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \mathbf{1}\left[\Omega_\tau = s\right] \quad a.s. \\ &= \frac{p(i_1)\cdots p(i_M)}{\prod_{k=1}^{M-1}(1 - \sum_{j=1}^{k} p(i_j))} \end{aligned} \tag{1.47}$$

for every $s = (i_1, \ldots, i_M)$ in $\Lambda(M; \mathcal{N})$. Consequently, the limit (1.9) exists for each $s = \{i_1, \ldots, i_M\}$ in $\Lambda^\star(M; \mathcal{N})$ as

$$Q_{\mathrm{LRU}}(s; \boldsymbol{p}) = \sum_{(j_1,\ldots,j_M)\in\Lambda(s|M;\mathcal{N})} \frac{p(j_1)\cdots p(j_M)}{\prod_{k=1}^{M-1}(1 - \sum_{\ell=1}^{k} p(j_\ell))} \tag{1.48}$$

where $\Lambda(s|M; \mathcal{N})$ is as defined in Section 1.10.1.

The miss rate of the LRU policy under IRM can then be evaluated from (1.11) as

$$M_{\mathrm{LRU}}(\boldsymbol{p}) = \sum_{(i_1,\ldots,i_M)\in\Lambda(M;\mathcal{N})} \frac{p(i_1)\cdots p(i_M)\left(1 - \sum_{j=1}^{M} p(i_j)\right)}{\prod_{k=1}^{M-1}(1 - \sum_{j=1}^{k} p(i_j))}. \tag{1.49}$$

13.1 A counterexample

Contrary to what transpired with RORA policies, the miss rate under the LRU policy is *not* Schur-concave in general, and consequently the folk theorem (1.3) does not hold. This is demonstrated through the following example developed for $M = 3$, $N = 4$, and the family of pmfs

$$\boldsymbol{p}(x, y) = (x, 1 - 2y - x, y, y), \quad 0 < y < \frac{1}{4}$$

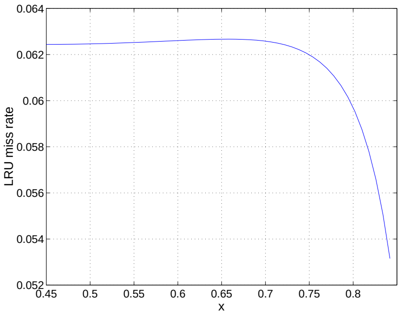

Figure 1.1. LRU miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.05$, $p(1) = x$ and $p(2) = 0.9 - p(1)$

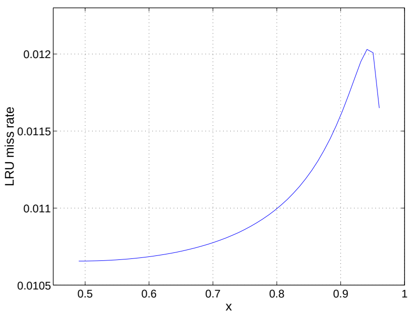

Figure 1.2. LRU miss rate when $M = 3$, $N = 4$, $y = p(3) = p(4) = 0.01$, $p(1) = x$ and $p(2) = 0.98 - p(1)$

with x in the interval $[\frac{1}{2} - y, 1 - 3y]$. Under these constraints, the components of the pmf $\boldsymbol{p}(x, y)$ are listed in decreasing order and for any given y it holds that $\boldsymbol{p}(x, y) \prec \boldsymbol{p}(x', y)$ whenever $x < x'$ in the interval $[\frac{1}{2} - y, 1 - 3y]$. Therefore, if the miss rate under LRU was indeed a Schur-concave function in the popularity pmf, we would expect the functions $x \to M_{\mathrm{LRU}}(\boldsymbol{p}(x, y))$ to be monotone decreasing in x on the interval $[\frac{1}{2} - y, 1 - 3y]$.

Figures 1.1 and 1.2 display the numerical values of $M_{\mathrm{LRU}}(\boldsymbol{p}(x, y))$ as a function of x with $y = 0.05$ and $y = 0.01$, respectively; this was done by numerical evaluation of (1.49). In both cases, the miss rate of the LRU policy is *not* monotone decreasing in x on the range $[\frac{1}{2} - y, 1 - 3y]$, with the trend becoming more pronounced with decreasing y. In short, the miss rate is *not* Schur-concave under the LRU policy.

13.2 LRU with Zipf-like popularity pmfs

While the miss rate is *not* Schur-concave under the LRU policy, the desired monotonicity (1.3) is nevertheless true in an asymptotic sense when the popularity pmf is restricted to the class of Zipf-like pmfs.

Theorem 1.19 *Assume the input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. Then, there exists $\alpha^\star = \alpha^\star(M, N) > 0$ and $\Delta > 0$ such that $M_{\mathrm{LRU}}(\boldsymbol{p}_\beta) < M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha)$ whenever $\alpha^\star < \alpha$ and $\alpha + \Delta < \beta$.*

This result is a byproduct of the asymptotic equivalence

$$\lim_{\alpha \to \infty} \frac{M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha)}{(M+1)^{-\alpha}} = 2 \tag{1.50}$$

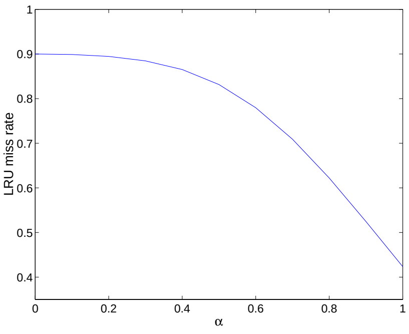

Figure 1.3. LRU miss rate when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for α small $(0 \leq \alpha \leq 1)$

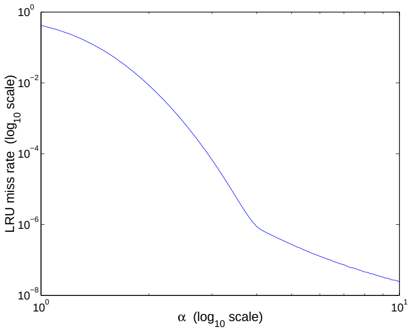

Figure 1.4. LRU miss rate when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for α large $(\alpha > 1)$

established in Vanichpun and Makowski (2004a). We have also carried out simulations of a cache operating under the LRU policy when the input has a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$.[8] The number of documents is set at $N = 1,000$ while the cache size is $M = 100$. The miss rate of the LRU policy is displayed in Figure 1.3 and 1.4 for α small $(0 \leq \alpha \leq 1)$ and α large $(\alpha > 1)$, respectively. It appears that the miss rate is indeed decreasing as the skewness parameter α increases across the *entire* range of α. This suggests that the folk theorem on miss rates probably holds for the LRU policy when the comparison is made within the class of Zipf-like popularity pmfs, hence the following

Conjecture 1.20 *For arbitrary cache size M and number N of documents, the function $\alpha \to M_{\mathrm{LRU}}(\boldsymbol{p}_\alpha)$ is strictly decreasing on $[0, \infty)$.*

14. The output under the LRU policy

With the expressions (1.47) for the stationary distribution of the LRU cache state, it is a simple matter to check for each $i = 1, \ldots, N$, that

$$\begin{aligned} m_{\mathrm{LRU}}(i;\boldsymbol{p}) &= \sum_{s \in \Lambda_i(M;\mathcal{N})} \pi_{\mathrm{LRU}}(s;\boldsymbol{p}) \\ &= \sum_{s \in \Lambda_i(M;\mathcal{N})} \frac{p(i_1)\cdots p(i_M)}{\prod_{k=1}^{M-1}(1 - \sum_{j=1}^{k} p(i_j))} \end{aligned} \tag{1.51}$$

[8] We choose simulations over numerical evaluation of (1.49) because this expression is not suitable for numerical evaluation due to a combinatorial explosion.

where $\Lambda_i(M;\mathcal{N})$ denote the set of elements in $\Lambda(M;\mathcal{N})$ which do not contain i, i.e., $\Lambda_i(M;\mathcal{N}) := \{s = (i_1, \dots i_M) \in \Lambda(M;\mathcal{N}) : \ i \notin s\}$. Theorem 1.2 then gives the output popularity pmf in the form

$$p^\star_{\mathrm{LRU}}(i) = \frac{p(i)}{M_{\mathrm{LRU}}(\boldsymbol{p})} \sum_{s\in\Lambda_i(M;\mathcal{N})} \frac{p(i_1)\cdots p(i_M)}{\prod_{k=1}^{M-1}(1-\sum_{j=1}^{k} p(i_j))} \tag{1.52}$$

for each $i = 1, \dots, N$, as we make use of (1.16). We begin with a positive result.

LEMMA 1.21 *The LRU policy is a good policy.*

In what follows, let $\boldsymbol{p}^\star_\alpha$ denote the popularity pmf of the output induced by an input with Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ (instead of the more cumbersome $\boldsymbol{p}^\star_{\mathrm{LRU},\alpha}$).

14.1 Another counterexample

In view of Lemma 1.21, it is tempting to expect that the majorization comparison $\boldsymbol{p}^\star_{\mathrm{LRU}} \prec \boldsymbol{p}$ also holds under the LRU policy. This is not the case as the following example demonstrates: With $M = 3$ and $N = 4$ under the Zipf-like popularity pmf (1.46) with $\alpha = 3$, we have computed the output popularity pmf under the LRU policy using (1.52). The numerical values of both input and output popularity pmfs are presented in Table 1.1.

Table 1.1. $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}^\star_\alpha$ under the LRU policy when the input distribution is Zipf-like with parameter $\alpha = 3$

i	*1*	*2*	*3*	*4*
$\boldsymbol{p}_\alpha$	0.8491	0.1061	0.0314	0.0133
$\boldsymbol{p}^\star_\alpha$	0.0118	0.2031	0.3853	0.3998

By the definition of majorization (1.17)-(1.18), the comparison $\boldsymbol{p}^\star_\alpha \prec \boldsymbol{p}_\alpha$ requires

$$\min_{i=1,\dots,N} p_\alpha(i) \le \min_{i=1,\dots,N} p^\star_\alpha(i), \tag{1.53}$$

in clear contradiction with Table 1.1, and therefore does not hold. On the other hand, the comparison $\boldsymbol{p}_\alpha \prec \boldsymbol{p}^\star_\alpha$ is not valid either since it calls for the unmet requirement

$$\max_{i=1,\dots,N} p_\alpha(i) \le \max_{i=1,\dots,N} p^\star_\alpha(i). \tag{1.54}$$

In short, $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}_\alpha^\star$ are not comparable in the majorization ordering. This situation does not represent an isolated incident as the next theorem shows; its proof is available in Vanichpun and Makowski (2004b).

Theorem 1.22 *Assume the input to have a Zipf-like popularity pmf $\boldsymbol{p}_\alpha$ for some $\alpha \geq 0$. If the number of documents N and the cache size M satisfy the condition $N < M!$, then under the LRU policy, there exists $\alpha^\star = \alpha^\star(M, N)$ such that $\boldsymbol{p}_\alpha^\star \prec \boldsymbol{p}_\alpha$ does not hold whenever $\alpha > \alpha^\star$.*

14.2 A conjecture

Theorems 1.15 and 1.17 were valid for *all* values of M and N, and for *arbitrary* admissible pmfs. While the counterexamples discussed earlier dash our hope to get an analogous result for the LRU policy, the possibility remains, fueled by Corollary 1.16, that the positive result is nevertheless valid in some appropriate range of the parameters M and N. We now explore this issue still with Zipf-like popularity pmfs (1.46).

Conjecture 1.23 *Assume that the popularity pmf is the Zipf-like pmf (1.46) with $\alpha \geq 0$. For each $N = 1, 2, \ldots$, there exists an integer $M^\star = M^\star(\alpha; N)$ with $1 \leq M^\star < N$ such that $\boldsymbol{p}_\alpha^\star \prec \boldsymbol{p}_\alpha$ under the LRU policy whenever $M = 1, \ldots, M^\star$.*

In support of this conjecture, we have carried out simulations of the cache operating under the LRU policy when the input pmf is Zipf-like with parameter $\alpha = 0.8, 1$ and 2 and with $N = 1,000$. We find the output popularity pmfs for different values of cache size, namely $M = 10, 50, 100, 500$. The resulting output popularity pmfs in the original order of documents are shown in Figure 1.5, while the results after rearranging documents in the decreasing order of their output probabilities are displayed in Figure 1.6.

From Figure 1.6 (a), when $\alpha = 0.8$, the comparison $\boldsymbol{p}_\alpha^\star \prec \boldsymbol{p}_\alpha$ holds for $M = 10, 50$. Indeed, from their respective plots, we observe that the pmfs $\boldsymbol{p}_\alpha$ and $\boldsymbol{p}_\alpha^\star$ when arranged in decreasing order intersect only once, namely $p_\alpha^\star([i]) \leq p_\alpha(i)$, $i = 1, \ldots, k$, and $p_\alpha^\star([i]) \geq p_\alpha(i)$, $i = k+1, \ldots, N$, for some $k = 1, \ldots, N-1$, where $p_\alpha^\star([1]) \geq p_\alpha^\star([2]) \geq \ldots \geq p_\alpha^\star([N])$ are the components of $\boldsymbol{p}_\alpha^\star$ arranged in decreasing order. This is the sufficient condition for majorization comparison provided in Proposition 1.3.

However, for $\alpha = 0.8$ and $M = 100, 500$, despite the fact that in Figure 1.6 (a), $\boldsymbol{p}_\alpha^\star$ looks uniform in the range where document rank is smaller than M, the comparison $\boldsymbol{p}_\alpha^\star \prec \boldsymbol{p}_\alpha$ is invalid since the necessary condition (1.53) does not hold. This violation, $\min_{i=1,\ldots,N} p_\alpha^\star(i) < p_\alpha(N)$, can be easily seen from Figure 1.5 (a) or from the subplot inside Figure 1.6 (a). For $\alpha = 1$ and 2, by the same arguments, we conclude from Figures 1.5

(b)-(c) and 1.6 (b)-(c) that the comparison $\boldsymbol{p}^\star_\alpha \prec \boldsymbol{p}_\alpha$ holds for $M = 10$ but does not hold for other cache sizes $M = 50, 100, 500$. Hence, these experimental results agree with Conjecture 1.23, and suggest that the value of $M^\star(\alpha; N)$ in Conjecture 1.23 decreases as α increases.

Acknowledgments

This material is based upon work supported by the Space and Naval Warfare Systems Center – San Diego under Contract No. N66001-00-C-8063. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Space and Naval Warfare Systems Center – San Diego.

References

V. Almeida, A. Bestavros, M. Crovella and A. de Oliveira. Characterizing reference locality in the Web. In *Proceedings of PDIS'96*, pp. 92-107, Miami (FL), December 1996.

O.I. Aven, E.G. Coffman and Y.A. Kogan. *Stochastic Analysis of Computer Storage*. D. Reidel Publishing Company, Dordrecht (Holland), 1987.

F. Baccelli and A.M. Makowski. Queueing models for systems with synchronization constraints. *Proceedings of the IEEE*, 77:138–161, 1989.

L.A. Belady. A study of replacement algorithms for a virtual-storage computer. *IBM Systems Journal*, 5:78–101, 1966.

L. Breslau, P. Cao, L. Fan, G. Phillips and S. Shenker. Web caching and Zipf-like distributions: Evidence and implications. In *Proceedings of IEEE INFOCOM 1999*, New York (NY), March 1999.

E. Coffman and P. Denning. *Operating Systems Theory*, Prentice-Hall, Englewood Cliffs (NJ), 1973.

P.J. Denning. The working set model for program behavior. *Communications of the ACM*, 11:323–333, 1968.

R. Fonseca, V. Almeida, M. Crovella and B. Abrahao. On the intrinsic locality of Web reference streams. In *Proceedings of IEEE INFOCOM 2003*, San Francisco (CA), April 2003.

S. Gadde, J.S. Chase and M. Rabinovich. Web caching and content distribution: A view from the interior. *Computer Communications*, 24:222–231, 2001.

E. Gelenbe. A unified approach to the evaluation of a class of replacement algorithms. *IEEE Transactions on Computers*, 22:611–618, 1973.

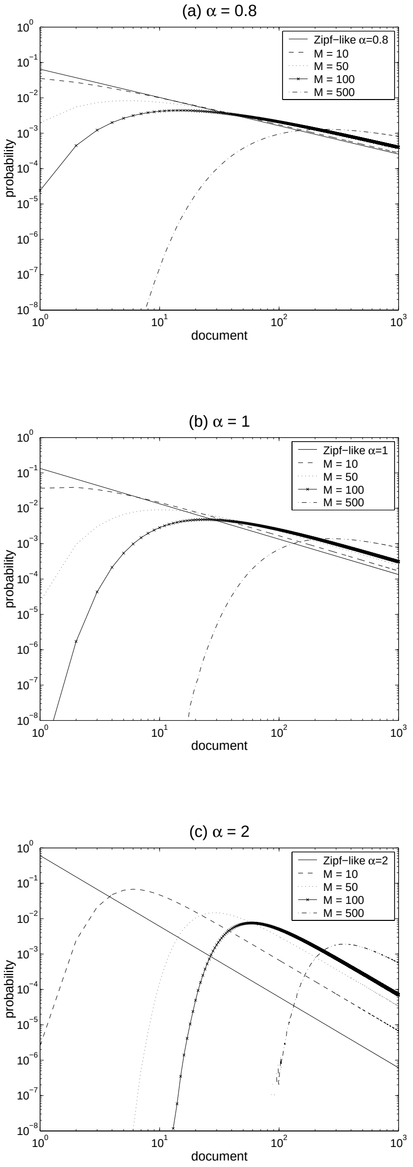

Figure 1.5. LRU output popularity pmf with different cache sizes M when the input has a Zipf-like pmf with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$

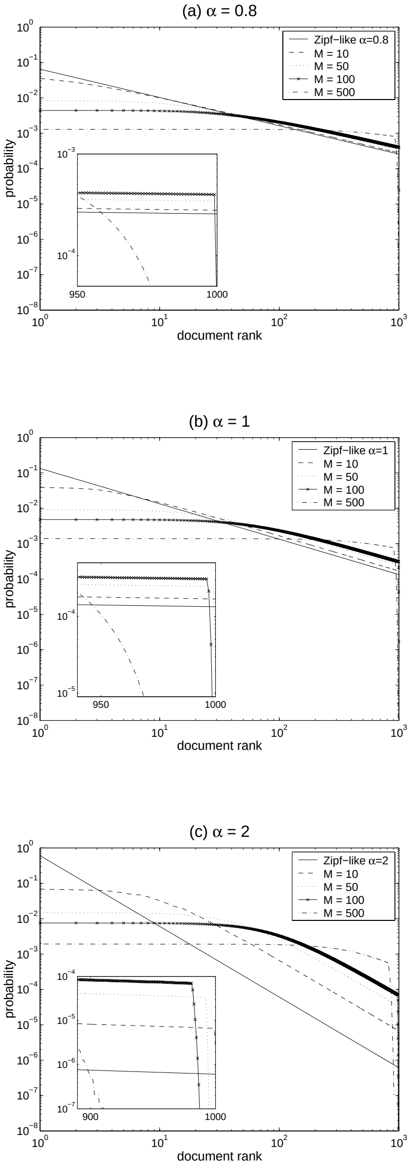

Figure 1.6. LRU output popularity pmf with different cache sizes M when the input has a Zipf-like pmf with (a) $\alpha = 0.8$, (b) $\alpha = 1$ and (c) $\alpha = 2$. Documents are ranked according to their probabilities.

P.R. Jelenkovic and A. Radovanovic. Asymptotic insensitivity of least-recently-used caching to statistical dependency. In *Proceedings of IEEE INFOCOM 2003*, San Francisco (CA), April 2003.

S. Jin and A. Bestavros (2000a). GreedyDual* Web caching algorithm: Exploiting the two sources of temporal locality in Web request streams. In *Proceedings of the 5th International Web Caching and Content Delivery Workshop*, Lisbon, Portugal, May 2000.

S. Jin and A. Bestavros (2000b). Sources and characteristics of Web temporal locality. In *Proceedings of MASCOTS'2000*, San Francisco (CA), August 2000.

A. Mahanti, C. Williamson and D. Eager. Temporal locality and its impact on Web proxy cache performance. *Performance Evaluation*, Special Issue on Internet Performance Modelling, 42:187–203, 2000.

A.W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York (NY), 1979.

V. Phalke and B. Gopinath. An inter-reference gap model for temporal locality in program behavior. In *Proceedings of ACM SIGMETRICS 1995*, pp. 291–300, May 1995, Ottawa, Ontario, Canada.

S. Vanichpun. *Comparing Strength of Locality of Reference – Popularity, Majorization, and Some Folk Theorems for the Miss Rates and the Output of Caches*. Ph.D. Dissertation, Department of Electrical and Computer Engineering, University of Maryland, College Park (MD), Expected December 2004.

S. Vanichpun and A.M. Makowski (2002). The effects of positive correlations on buffer occupancy: Lower bounds via supermodular ordering. in *Proceedings of IEEE INFOCOM 2002*, New York (NY), June 2002.

S. Vanichpun and A.M. Makowski (2004a). Comparing strength of locality of reference – Popularity, majorization, and some folk theorems. in *Proceedings of IEEE INFOCOM 2004*, Hong Kong, March 2004.

S. Vanichpun and A.M. Makowski (2004b). The output of a cache under the independent reference model – Where did the locality of reference go? In *Proceedings of ACM Sigmetrics – Performance 2004*, New York (NY), June 2004.

J. Wang (1999). *A Survey of Web Caching Schemes for the Internet.* Technical Report TR99-1747, Department of Computer Science, Cornell University, Ithaca (NY), May 1999.